Polygraph

VOLUME 28 1999 NUMBER 1

Special Edition

Chart Interpretation

Published Quarterly

P.O. Box 8037, Chattanooga, Tennessee 37414-0037

The Utah Numerical Scoring System

Brian G. Bell[1], David C. Raskin[2], Charles R. Honts[3], & John C. Kircher[1]

Abstract

The Utah method for numerically evaluating polygraph charts is a highly reliable and valid method for scoring specific-incident, comparison-question tests. For respiration, electrodermal activity (skin conductance or skin resistance), relative blood pressure (cardiograph), and peripheral vasomotor activity (finger plethysmograph), a score from +3 to -3 is assigned for each presentation of a relevant question. The reaction to the relevant question is compared to the reaction to a nearby comparison (control) question. A positive score is assigned when the psychophysiological reaction is greater to the comparison question than to the relevant question, a negative score is assigned when the reaction is greater to the relevant question, and a zero is assigned when the responses to the relevant and comparison questions are approximately equal. Scores are based on the criteria described in the present report. Common artifacts that may affect numerical evaluations are discussed, as are limitations of this scoring system.

Key words: comparison-question tests, detection of deception, numerical scoring, polygraph

This report describes the Utah method for numerically evaluating polygraph charts from specific-incident, comparison-question tests. The development of the Utah method was preceded by numerical scoring techniques introduced by Backster (1969) and the U.S. Army (Weaver, 1980; 1985). Although these early scoring systems represented major improvements over global approaches to chart evaluation, many of their scoring rules had no scientific basis and had not been validated by scientific research. The Backster system in particular had been shown to be biased against truthful subjects, (Raskin, 1986) and consisted of many complex scoring rules that made it difficult to evaluate polygraph charts reliably. The Utah system was developed to simplify the scoring process, reduce bias, and improve the accuracy of decisions. It consists of relatively few rules that may be applied with considerable consistency by different numerical evaluators after a brief period of training.

The reliability of the Utah scoring system has been evaluated in several laboratory experiments at the University of Utah. The results of five such studies are summarized in Table 1. On average, the interrater reliability of the Utah system exceeded .90, as measured by the correlations between total numerical scores assigned by two or more evaluators. The percent agreement on decisions exceeded 95% when both numerical evaluators reached a definite decision. Similar reliabilities between raters who use the Utah system have been obtained

Acknowledgements

Portions of the research described in this report were funded by a Marriner S. Eccles Graduate Fellowship in Political Economy from the University of Utah, the National Institute of Law Enforcement and Criminal Justice (Contract 75-NI-99-0001), and the National Institute of Justice (Grants 8l IJ-CX-0051 and 85-IJ-CX-0040). The views expressed in this article are those of the authors and do not reflect the official policy or position of the National Institute of Justice or the U.S. Government. The authors also gratefully acknowledge comments by Paul Bernhardt on an early draft of the manuscript. For reprints, contact Dr. John Kircher, Department of Educational Psychology, 327 MBH, University of Utah, Salt Lake City, UT 84112.

[1] Department of Educational Psychology, 327 MBH, University of Utah, Salt Lake City, UT 84112.
[2] P.O. Box 2419, Homer, AK 99603.
[3] Department of Psychology, Boise State University, Boise, ID 83725.

in field studies. For example, the interrater correlation was .94 in a field study by Honts and Raskin (1988). These reliabilities far exceed standards of acceptable interrater reliability for psychological tests as established by the American Psychological Association (1985).

Table 1. Reliability of the Utah System of Numerical Scoring in Laboratory Studies

Study	Agreement on Decisions Between Original Examiner and Independent Evaluator*	Correlation Between Numerical Scores of Independent Evaluator and Original Examiner
Podlesny & Raskin (1978)	100%	.97
Rovner et al. (1979)	95%	.97
Kircher & Raskin (1988)	99%	.97
Honts et al. (1994)	96%	.92
Horowitz et al. (1997)	.98**	.92

* Includes only cases in which both examiners made a decision (excludes inconclusives)

** Only Kappa was reported in this study.

The validity of the Utah system of numerical evaluation has also been established. Table 2 presents decision accuracies from several laboratory experiments. Excluding inconclusive outcomes, the overall percentage of correct decisions was 91% for guilty subjects and was 89% for innocent subjects.

The results from field studies with the Utah system are consistent with those reported in Table 2 (Honts & Raskin, 1988; Raskin, 1976; Raskin, Kircher, Honts, & Horowitz, 1988). In one field study, two numerical evaluators independently evaluated the polygraph charts using the Utah system (Raskin, 1976). Their decisions were 100% correct for both guilty and innocent suspects. In another study, decisions were 92% correct for guilty suspects and 100% correct for innocent suspects (Honts & Raskin, 1988).

Overview of the Utah Scoring System

The Utah scoring system, when used with the probable-lie and directed-lie comparison question tests, assigns numerical scores by assessing differences between relevant and comparison questions. Scores are assigned on a 7-point scale that ranges from -3 to +3. The reaction to a relevant question is compared to the reaction produced by a temporally adjacent, comparison question. If a relevant question was presented between two comparison questions, its reaction is compared to the reaction to the comparison question that produced the stronger physiological response.

For each channel, the relative size of the reactions to the comparison and relevant questions is evaluated and quantified. Positive scores are assigned when the physiological reaction to the comparison question was greater than the reaction to the relevant question. Negative scores are assigned when the reaction to the relevant question was greater, and zero is assigned when reactions to the relevant and comparison questions are not noticeably different. In general, a noticeable difference between the reactions to the comparison and relevant questions is assigned a score of 1. A strong, clear difference between

Table 2. Validity of the Utah System of Numerical Scoring in Laboratory Studies

	Guilty					Innocent				
Study	N	Correct	Incorrect	Inconclusive	% Correct*	N	Correct	Incorrect	Inconclusive	% Correct*
Raskin & Hare (1978)	24	88%	0%	12%	100%	24	75%	4%	21%	95%
Podlesny & Raskin (1978)	20	70%	15%	15%	82%	20	90%	5%	5%	95%
Rovner et al. (1979)[a]	24	88%	0%	12%	100%	24	88%	8%	4%	92%
Kircher & Raskin (1988)	50	88%	6%	6%	94%	50	86%	6%	8%	93%
Honts et al. (1994)	20	70%	20%	10%	78%	20	75%	10%	15%	88%
Horowitz et al. (1997)[b]	15	53%	20%	27%	73%	15	80%	13%	7%	86%

* The percent correct was calculated by dividing the number correct by the sum of the number correct and the number incorrect.

[a] Excludes 24 countermeasure-trained subjects.
[b] Excludes 90 subjects given relevant-irrelevant and directed lie tests.

the reactions is assigned a 2. A score of 3 is assigned when there is a dramatic difference between the reactions to the two questions, the tracing is stable, and the stronger response is the largest on the chart for that physiological measure.

In a single-issue test, all relevant questions must be answered truthfully or all must be answered deceptively. In this case, the scores for all presentations of relevant questions are summed. The subject is reported as deceptive if the total score is -6 or lower, truthful if the total is +6 or higher, and inconclusive if the total is between -6 and +6.

For mixed issue tests, such as the Modified General Question Test, some relevant questions can be answered truthfully while others can be answered deceptively. In this case, a separate total is obtained for each relevant question. When the total score for a single relevant question is -3 or lower, the subject's answer to that question is considered deceptive. When the total score for a single relevant question is +3 or higher, the subject's answer is considered truthful. When the total score for a question is between -3 and +3, the outcome is considered inconclusive. However, if the total score for all relevant questions combined is at least +6 or -6, and the total scores for each relevant question are in the same direction (all positive or all negative), the subject is considered truthful or deceptive, respectively, to each relevant question.

Scoring Criteria

A total of ten scoring criteria are used to assess the relative strength of physiological reactions to relevant and comparison questions. The criteria change depending on the physiological measure being evaluated. Scores are assigned to respiration, electrodermal, cardiograph, and finger plethysmograph channels.

Respiration

For a given relevant question, changes in respiration are evaluated first because deep breaths may affect how other channels are evaluated. In general, a reaction to a question is indicated by suppressed respiratory activity. The greater the suppression, the stronger the reaction. Suppression is indicated primarily by a reduction in the amplitude of at least two successive respiration cycles following question onset and brief periods of apnea (cessation of breathing). A rise in the respiration baseline, as indicated by a rise in the bottoms of at least two respiration cycles, is another criterion for scoring a reaction. An increase in cycle time (slowing of respiration rate) is also a criterion but is less heavily weighted than changes in amplitude, apnea, and baseline increase. Increases in respiratory activity, such as increased amplitude, speeding of respiration, and drops in respiration baseline, are not indications of a reaction and are not criteria for scoring.

Although thoracic and abdominal respiration are recorded on separate channels of the polygraph, only one numerical score is assigned that is based on a composite of both channels. Respiration reactions to the comparison and relevant questions are evaluated by noting the combined amount of reaction in both respiration channels for the relevant question and for the comparison question. A single numerical score is then assigned based on the difference in the composite reactions to the relevant and comparison questions. Thus, the numerical score for one relevant question may be based on observed changes in thoracic respiration, abdominal respiration, or a composite of both, depending on the total amount of change observed in the two channels.

Electrodermal Activity

The electrodermal channel is evaluated next. Numerical scores for electrodermal activity are based mainly on changes in peak amplitude. The amplitude of a reaction is defined as the greatest difference between any low point and subsequent high point that occurs within the scoring window (described below). Amplitude may be measured by using the numerical scoring subprogram in the Stoelting Computerized Polygraph System (CPS) or a similar system. If only printed or inked charts are available, the amplitude is measured with a ruler to the nearest 0.5 millimeter. For each relevant question, a score of 1 is assigned if the amplitude of the reaction to the relevant or comparison question is twice as large as the amplitude of the reaction to the other question. A score of 2 is assigned when the amplitude of the reaction is three times as

large, and a score of 3 is assigned when the amplitude is four times as large. However, a score of 3 may be assigned only when the baseline is stable and the reaction is the largest on the chart. The baseline is considered unstable if there are many nonspecific electrodermal responses on the chart. Under those conditions, a score of 3 cannot be assigned.

The duration of the electrodermal reaction and its complexity (the number of waves or fluctuations that occur within the scoring window) are also considered when scoring the electrodermal channel. Reactions that have clearly longer duration or greater complexity may increase the score from 0 to 1 or from 1 to 2. The larger score may be assigned if the ratio of the amplitudes is at least 1.5:1 or 2.5:1, and the larger reaction has longer duration and/or is more complex. However, a score of 1 or 2 cannot be assigned if the reactions differ only in duration and/or complexity, and these criteria are not used to assign a score of 3.

Cardiograph

For the cardiograph channel, reactions are measured as rises in the baseline. The numerical score is based primarily on the largest rise in the baseline that occurs within the scoring window. Again, use of a computer-scoring algorithm, such as the CPS, or a ruler is recommended for making measurements of increases in the baseline. A minimum ratio of 1.5 to 1 is required for a score of 1. Measurements of baseline increases are made on the diastolic side of the waveform because the diastolic points show greater change than the systolic points and are easier to see. However, increases in the systolic points may be used if it is unclear whether to assign a 0 or 1 or to assign a 1 or 2 based on the diastolic points. The duration of the response is also considered. The rules described above for electrodermal reactions apply to the cardiograph; reactions with longer duration may increase the numerical score from 0 to 1 or from 1 to 2.

Finger Plethysmograph

Peripheral vasomotor activity is measured from a photoplethysmograph attached to the tip of the finger. Constriction of blood vessels in the finger produces a reduction (constriction) in the amplitude of finger pulses. Numerical scores are based on the duration and magnitude of reductions in finger pulse amplitude. Responses of longer duration and/or magnitude are assigned larger numerical scores. Unlike the cardiograph and electrodermal channels, scores of 1 or 2 may be assigned to this response system when there is little or no difference in the reduction of pulse amplitude, but there is a clear difference in the duration of the reactions.

Scoring Windows

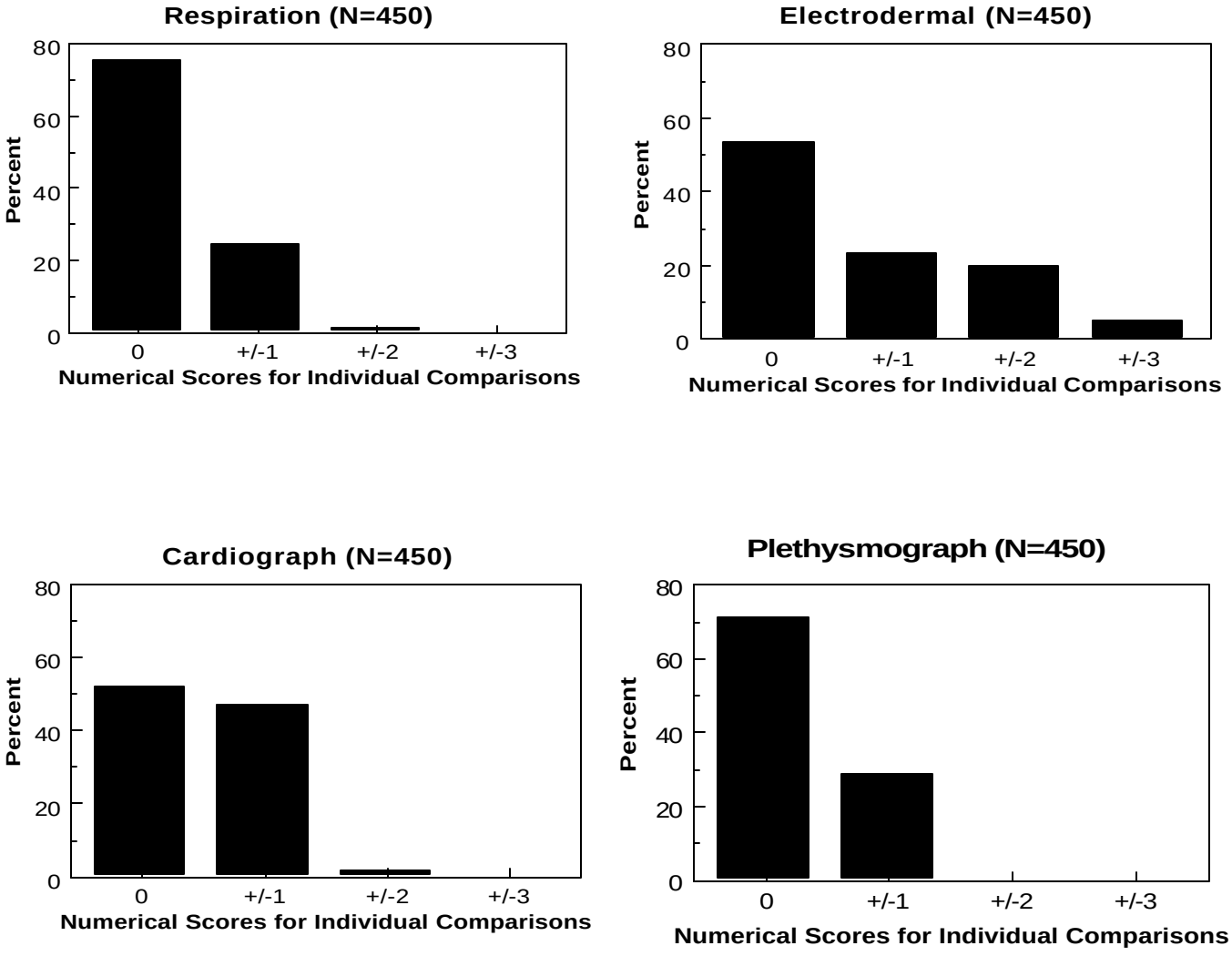
For all channels, the response is not scored unless it begins after question onset. However, the minimum latency for a response varies depending on the physiological measure. Respiratory and cardiograph reactions may be scored if they begin immediately after question onset. Electrodermal reactions are scored only if they begin at least 0.5 seconds after question onset, and finger pulse reactions are scored only if they begin at least 2 seconds after question onset. If a reaction begins prior to the minimum latency, a point of inflection or clear increase in slope that occurs after the minimum latency may be considered the beginning of the reaction. For all physiological measures, the reaction must begin no later than 5 seconds after the subject's answer, unless the subject characteristically has reactions that begin 5 to 8 seconds after answering. Such delayed reactions should be scored conservatively. Reactions that begin outside these scoring windows are not scored. The duration of a reaction that begins within the scoring window may be considered up to 20 seconds following question onset.

Distributions of Numerical Scores

We measured the frequency with which we observed physiological changes that met the criteria for 'noticeable', 'clear', and 'dramatic' differences in a random sample of 25 innocent and 25 guilty subjects who participated in a previous experiment (Kircher & Raskin, 1988). Guilty subjects had committed a mock theft, innocent subject did not commit the theft, and all subjects had been promised a substantial monetary bonus if they could convince the polygraph examiner that they were innocent of the crime. The

charts were scored by the second author (DCR) who had no contact with the subjects and was unaware of the subjects' guilt or innocence. The overall accuracy of decisions was 95% for guilty subjects and 96% for innocent subjects. For each physiological component, each of three relevant questions was scored against the probable-lie comparison question that immediately preceded it on each of three charts. This provided a sample of 450 numerical scores for each physiological component (3 relevant questions X 3 charts X 50 subjects). The absolute values of these scores are presented in Figure 1 for each physiological measure.

Figure 1.
Distributions of numerical scores for respiration, electrodermal, cardiograph, and finger plethysmograph.



As shown in Figure 1, numerical scores of '0' were assigned considerably more often than any other value. It also may be seen that the frequency with which different numerical scores were assigned depended on the physiological measure. More than 70% of the numerical scores were '0' for the respiration and plethysmograph channels, whereas approximately 50% of the scores were '0' for skin conductance and cardiograph channels. These results suggest that, on average, decisions will be based largely on the numerical scores assigned to the electrodermal and cardiograph channels.

Although electrodermal and cardiograph channels had about the same number of scorable (nonzero) differences, numerical scores of 2 or 3 were more common for the electrodermal channel. Since more scores of 2 and 3 were assigned to the electrodermal channel, it had more influence than the cardiograph on the total numerical scores and the outcomes of the polygraph tests. Together, the data indicate that the Utah scoring rules give greater weight to electrodermal reactions than to cardiovascular, respiration, or plethysmograph reactions. The relative weights given the four physiological measures by the Utah scoring rules are remarkably consistent with the optimal combinations of weighted physiological measures that are generated by our Computerized Polygraph System (CPS) (Kircher & Raskin, 1991).

It should be noted that scores of 3 were extremely rare. The rules allow for the assignment of scores of 3 to any channel, but in this sample of 450 comparisons, not a single score of 3 was assigned to respiration, cardiograph, or plethysmograph channels.

Artifacts

The quality of the tracing is considered when assigning scores. Artifacts, such as deep breaths, coughs, movements, and physiological abnormalities, affect how scores are assigned. To minimize the occurrence of artifacts, we instruct examinees to avoid movements and to breathe normally while the recordings are made. Nevertheless, artifacts may render the tracings unscorable.

If a deep breath occurs shortly before question onset, respiration should not be scored. If the deep breath is accompanied by physiological changes in other channels, the other channels may or may not be scored. If the reaction in the other channel began before the deep breath, then the portion preceding the deep breath may be used in scoring if it is larger than the reaction to the question to which it is compared. If it is smaller and is to a comparison question, then another comparison question may be used. The evaluator should also examine all of the charts for the subject and locate any other places where a deep breath occurred, especially at points where no question had been asked. If there is a similar physiological change at this point, then the reaction following the deep breath must not be used for scoring. If there is no reaction following the deep breath, then the reaction may be scored, but it should be scored conservatively.

If movements distort more than two successive pulses in the cardiovascular channels after question onset, the cardiovascular changes that occur after the movement should not be scored. If there is a reaction that precedes the artifact, it may be used for scoring if it is larger than the reaction to which it is compared. However, if only one or two pulses are distorted, it is usually possible to visually interpolate across the artifact and infer what the reaction would have been if the movement had not occurred. If multiple artifacts occur within the scoring window, it is usually not possible to score the response.

Physiological abnormalities, such as premature ventricular contractions (PVCs), may also render the cardiovascular reaction unscorable if they occur in the scoring window. PVCs are contractions of the left ventricle that occur before the left atrium has contracted and filled the left ventricle, causing very little blood to be pumped into the aorta. This is followed by a relatively long pause before the next ventricular contraction. During this pause, the drop in blood pressure produces a distinct downward deflection in the cardiovascular tracing. If two or three PVCs occur within the scoring window, the signal is usually so distorted that it is not possible to score it. However, a cardiovascular reaction

that occurred before the PVC may be scored. It is usually also possible to score the reaction if it contains only one PVC, although the subsequent rise in the tracing that is the recovery from the PVC should not be scored as a reaction.

Limitations

The research that supports the use of the Utah system of numerical scoring has been limited to specific-incident examinations. It has not been validated for employment screening or periodic testing of employees with access to sensitive information. Furthermore, most of our research has focused on the probable-lie test. Use of the numerical scoring system with the directed-lie has also been validated (Honts & Raskin, 1988; Horowitz et al., 1997), but the relevant-irrelevant and other types of tests have received almost no attention.

Most of the laboratory research with the Utah system has used a single-issue test that contains three repetitions of neutral, comparison, and relevant questions in the question sequence. Other question sequences or mixed-issue tests have not been tested extensively in our laboratory, although two field studies included at least one numerical evaluator who used the Utah system with the Modified General Question Test format (Raskin, 1976; Raskin et. al., 1988), and the accuracy of those decisions was comparable to those we have observed in our laboratory experiments. Therefore, there is evidence that the validity of the Utah scoring technique generalizes across similar test formats.

References

American Psychological Association (1985). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

Backster, C. (1969). *Technique fundamentals of the tri-zone polygraph test*. New York: Backster Research Foundation.

Honts, C.R., & Raskin, D.C. (1988). A field study of the validity of the directed lie control question. *Journal of Police Science and Administration*, 16, 56-61.

Honts, C. R., Raskin, D. C. & Kircher, J. C. (1994). Mental and physical countermeasures reduce the accuracy of polygraph tests. *Journal of Applied Psychology*, 79, 252-259.

Horowitz, S. W., Kircher, J. C., Honts, C. R., & Raskin, D. C. (1997). The role of comparison questions in physiological detection of deception. *Psychophysiology*, 34, 108-115.

Kircher, J. C. & Raskin, D. C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology*, 73, 291-302.

Kircher, J. C. & Raskin, D. C. (1991-98). Computerized Polygraph System (CPS) Versions 1.00-2.20. Scientific Assessment Technologies, Inc., 2532 Chadwick Street, Salt Lake City, UT 84106

Podlesny, J. A. & Raskin, D. C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology*, 15, 344-359.

Raskin, D. C., (1976). Reliability of chart interpretation and sources of errors in polygraph examinations. Report to the National Institute of Law Enforcement and Criminal Justice (Contract 75-NI-99-0001). Salt Lake City: University of Utah, Department of Psychology.

Raskin, D. C. & Hare, R. D. (1978). Psychopathy and detection of deception in a prison population. *Psychophysiology*, 15, 126-136.

Raskin, D. C. (1986). The polygraph in 1986: Scientific, political, and legal issues surrounding application and acceptance of polygraph evidence. *Utah Law Review*, 1986, 29-74.

Raskin, D. C., Kircher, J. C., Honts, C. R., & Horowitz, S. W. (1988). A study of the validity of polygraph examinations in criminal investigation. Final report to the National Institute of Justice (Grant No. 85-IJ-CX-0040). Salt Lake City: University of Utah, Department of Psychology.

Rovner, L. I., Raskin, D.C., & Kircher, J. C. (1979). Effects of information and practice on detection of deception. *Psychophysiology*, 16, 197-198. (Abstract).

Weaver, R. S. (1980). The numerical evaluation of polygraph charts: Evolution and comparison of three major systems. *Polygraph*, 9, 94-108.

Weaver, R. S. (1985). Effects of differing numerical chart evaluation systems on polygraph examination results. *Polygraph*, 14, 34-41.

Table 2. Validity of the Utah System of Numerical Scoring in Laboratory Studies

	Guilty					Innocent				
Study	N	Correct	Incorrect	Inconclusive	% Correct*	N	Correct	Incorrect	Inconclusive	% Correct*
Raskin & Hare (1978)	24	88%	0%	12%	100%	24	75%	4%	21%	95%
Podlesny & Raskin (1978)	20	70%	15%	15%	82%	20	90%	5%	5%	95%
Rovner et al. (1979)[a]	24	88%	0%	12%	100%	24	88%	8%	4%	92%
Kircher & Raskin (1988)	50	88%	6%	6%	94%	50	86%	6%	8%	93%
Honts et al. (1994)	20	70%	20%	10%	78%	20	75%	10%	15%	88%
Horowitz et al. (1997)[b]	15	53%	20%	27%	73%	15	80%	13%	7%	86%

* The percent correct was calculated by dividing the number correct by the sum of the number correct and the number incorrect.

[a] Excludes 24 countermeasure-trained subjects.
[b] Excludes 90 subjects given relevant-irrelevant and directed lie tests.